Durably reducing conspiracy beliefs through dialogues with AI

Year: 2024 | Citations: 212 | Authors: Thomas H. Costello, Gordon Pennycook, David G. Rand

Abstract

Conspiracy theory beliefs are notoriously persistent. Influential hypotheses propose that they fulfill important psychological needs, thus resisting counterevidence. Yet previous failures in correcting conspiracy beliefs may be due to counterevidence being insufficiently compelling and tailored. To evaluate this possibility, we leveraged developments in generative artificial intelligence and engaged 2190 conspiracy believers in personalized evidence-based dialogues with GPT-4 Turbo. The intervention reduced conspiracy belief by ~20%. The effect remained 2 months later, generalized across a wide range of conspiracy theories, and occurred even among participants with deeply entrenched beliefs. Although the dialogues focused on a single conspiracy, they nonetheless diminished belief in unrelated conspiracies and shifted conspiracy-related behavioral intentions. These findings suggest that many conspiracy theory believers can revise their views if presented with sufficiently compelling evidence. Editor’s summary Beliefs in conspiracies that a US election was stolen incited an attempted insurrection on 6 January 2021. Another conspiracy alleging that Germany’s COVID-19 restrictions were motivated by nefarious intentions sparked violent protests at Berlin’s Reichstag parliament building in August 2020. Amid growing threats to democracy, Costello et al. investigated whether dialogs with a generative artificial intelligence (AI) interface could convince people to abandon their conspiratorial beliefs (see the Perspective by Bago and Bonnefon). Human participants described a conspiracy theory that they subscribed to, and the AI then engaged in persuasive arguments with them that refuted their beliefs with evidence. The AI chatbot’s ability to sustain tailored counterarguments and personalized in-depth conversations reduced their beliefs in conspiracies for months, challenging research suggesting that such beliefs are impervious to change. This intervention illustrates how deploying AI may mitigate conflicts and serve society. —Ekeoma Uzogara INTRODUCTION Widespread belief in unsubstantiated conspiracy theories is a major source of public concern and a focus of scholarly research. Despite often being quite implausible, many such conspiracies are widely believed. Prominent psychological theories propose that many people want to adopt conspiracy theories (to satisfy underlying psychic “needs” or motivations), and thus, believers cannot be convinced to abandon these unfounded and implausible beliefs using facts and counterevidence. Here, we question this conventional wisdom and ask whether it may be possible to talk people out of the conspiratorial “rabbit hole” with sufficiently compelling evidence. RATIONALE We hypothesized that interventions based on factual, corrective information may seem ineffective simply because they lack sufficient depth and personalization. To test this hypothesis, we leveraged advancements in large language models (LLMs), a form of artificial intelligence (AI) that has access to vast amounts of information and the ability to generate bespoke arguments. LLMs can thereby directly refute particular evidence each individual cites as supporting their conspiratorial beliefs. To do so, we developed a pipeline for conducting behavioral science research using real-time, personalized interactions between research subjects and AI. Across two experiments, 2190 Americans articulated—in their own words—a conspiracy theory in which they believe, along with the evidence they think supports this theory. They then engaged in a three-round conversation with the LLM GPT-4 Turbo, which we prompted to respond to this specific evidence while trying to reduce participants’ belief in the conspiracy theory (or, as a control condition, to converse with the AI about an unrelated topic). RESULTS The treatment reduced participants’ belief in their chosen conspiracy theory by 20% on average. This effect persisted undiminished for at least 2 months; was consistently observed across a wide range of conspiracy theories, from classic conspiracies involving the assassination of John F. Kennedy, aliens, and the illuminati, to those pertaining to topical events such as COVID-19 and the 2020 US presidential election; and occurred even for participants whose conspiracy beliefs were deeply entrenched and important to their identities. Notably, the AI did not reduce belief in true conspiracies. Furthermore, when a professional fact-checker evaluated a sample of 128 claims made by the AI, 99.2% were true, 0.8% were misleading, and none were false. The debunking also spilled over to reduce beliefs in unrelated conspiracies, indicating a general decrease in conspiratorial worldview, and increased intentions to rebut other

conspiracy believers. CONCLUSION Many people who strongly believe in seemingly fact-resistant conspiratorial beliefs can change their minds when presented with compelling evidence. From a theoretical perspective, this paints a surprisingly optimistic picture of human reasoning: Conspiratorial rabbit holes may indeed have an exit. Psychological needs and motivations do not inherently blind conspiracists to evidence—it simply takes the right evidence to reach them. Practically, by demonstrating the persuasive power of LLMs, our findings emphasize both the potential positive impacts of generative AI when deployed responsibly and the pressing importance of minimizing opportunities for this technology to be used irresponsibly. Dialogues with AI durably reduce conspiracy beliefs even among strong believers. (Left) Average belief in participant’s chosen conspiracy theory by condition (treatment, in which the AI attempted to refute the conspiracy theory, in red; control, in which the AI discussed an irrelevant topic, in blue) and time point for study 1. (Right) Change in belief in chosen conspiracy from before to after AI conversation, by condition and participant’s pretreatment belief in the conspiracy.